Predicting the Quality of Collaborative Problem Solving Through Linguistic Analysis of Discourse

Joseph M. Reilly
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA 02138
1 (617) 496-5164
josephreilly@fas.harvard.edu

Bertrand Schneider
Harvard Graduate School of Education
13 Appian Way
Cambridge, MA 02138
1 (617) 496-2094
bertrand_schneider@gse.harvard.edu

ABSTRACT

Collaborative problem solving in computer-supported environments is of critical importance to the modern workforce. Coworkers or collaborators must be able to co-create and navigate a shared problem space using discourse and non-verbal cues. Analyzing this discourse can give insights into how consensus is reached and can estimate the depth of their understanding of the problem. This study uses Coh-Metrix, a natural language processing tool that measures cohesion, to analyze participant discourse from a recent multi-modal learning analytics study where novice programmers collaborated to use a block-based programming language to instruct a robot on how to solve a series of mazes. We significantly correlated thirty-five Coh-Metrix indices from the transcripts of dyads' discourse with collaboration, learning gains, and multimodal sensor values. We then fit a variety of machine learning classifiers to predict collaboration using the indices generated by Coh-Metrix as features. This study paves the way for real-time detection of (un)productive interactions from multimodal data and could lead to real-time interventions to support collaborative learning.

Keywords

Collaboration, computer-supported collaborative work, multi-modal learning analytics, Coh-Metrix.

1. INTRODUCTION

Collaborative problem solving with computer-based or computer-supported environments has long been a focus of research on educational technologies [1] and is now seen as a 21st century learning objective of critical importance to the workforce [2]. Discourse of collaborators who co-create and navigate a shared problem space can give insights into how consensus is reached and the depth of their understanding. Because qualitative coding of transcripts or video is laborious and time-consuming, capturing and quantifying the quality of social interactions remains a challenge in the social sciences.

Joseph Reilly and Bertrand Schneider "Predicting the Quality of Collaborative Problem Solving Through Linguistic Analysis of Discourse" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 149 - 157

Applied natural language processing can automate the analysis of large corpora of human language and is a foundational technique of educational data mining. Coh-Metrix [3], a tool originally developed to measure text difficulty, has been used to evaluate online discussion transcripts and intelligent tutoring system dialogue. By applying the tool to discourse from a collaborative problem-solving activity, we hypothesize that certain markers can indicate the quality of the collaboration and could be used to predict how well groups work together from their speech patterns.

This paper uses Coh-Metrix to analyze participant discourse from a recent multi-modal learning analytics (MMLA) [7] study where novice programmers used a block-based programming language to program a robot to solve a series of mazes [4]. Preliminary results from this study indicate the importance of speech equity and talking time; however, the full transcripts of the discourse have not yet been analyzed. In this paper, we explore multiple indices from the transcripts that are correlated with collaboration, learning gains, and multimodal sensor values. We then explore ways of predicting collaboration using the indices generated by Coh-Metrix as features fed into a variety of machine learning classifiers. Finally, we discuss these results and conclude with future avenues for this research.

2. LITERATURE REVIEW

A few decades ago, computer-supported collaborative learning (CSCL) emerged as a statement against the over-individualization of educational technology, emphasizing that collaborative learning can be fostered by carefully designed computer-supported activities [5]. Collaboration analysis in computer-supported environments has explored what facets of collaborative processes are essential for successful problem-solving and learning [29] with a particular emphasis on what aspects of these analyses can be automated [30].When studying the process of collaboration, a 'Joint Problem Space' emerges that takes the form of a socially-negotiated knowledge structure that combines goal-setting, descriptions of the problem, and available actions [1]. This problem space can be understood through discourse analysis in conjunction with any other data collected on group behaviors.

In the search for new tools to analyze and model interactions between collaborators [6], Multi-Modal Learning Analytics (MMLA) has emerged. Sensors continue to get cheaper and easier to use while providing rich streams of data which can be used in conjunction to track and assess collaboration [7]. Data from multiple high frequency sensors can triangulate difficult to measure constructs and enhance overall predictive performance [31]. Analyzing features engineered from sensor data as well as dyadic

discourse provides a deeper view of the joint problem space, which includes nonverbal communication, posture, eye gaze, and arousal, among other possible metrics.

Coh-Metrix is an online tool that measures 106 linguistic features related to text easability, cohesion, lexical sophistication, syntactic complexity, and readability [3]. To differentiate between coherence and cohesion, the developers of this tool view cohesion as a quality of the text or discourse that can be directly measured, while coherence is in the mind of the reader [8]. The Coh-Metrix indices generally indicate the presence or absence of cohesive cues that tie the text together and make it easier or harder to understand. The tool focuses on local and overall text cohesion versus global cohesion [9].

Coh-Metrix has mainly been used in analyzing text readability and writing quality but has been applied successfully to other domains as well. Its indices have been used to detect lying in online discourse with one group member specified as sender and other as the receiver [10]. Tutor dialogue in AutoTutor was compared to naturalistic dialogue with a human tutor using Coh-Metrix to see how the dialogues differed on cohesion indices [11]. Indices for cohesion were also used to train affect detectors for AutoTutor users with the intention of developing real-time affect detectors based on cohesion [12]. The tool has also been used with online discussion transcripts to classify online discourse for levels of cognitive presence, and a classifier using Coh-Metrix features outperformed a similar algorithm using bag-of-words features [13].

Rarely do Coh-Metrix studies use transcripts of oral dialogue or assume participants are both novices (i.e. the task is not an expert-novice tutoring scenario.) Additionally, these indices have been sparingly used in MMLA research. In an MMLA study on a similar collaborative task, verbal coherence positively correlated with learning gains and significantly differed by condition [14]. Researchers then used language metrics to predict learning gains via support vector machine (SVM). Initial work from this current study has indicated that amount of talking and equity of talk time may be important indicators of good collaboration [15] but this discourse has not been analyzed in-depth yet.

3. RESEARCH QUESTIONS

This study attempts to answer the following research questions (RQs):

RQ1: Are Coh-Metrix indices derived from transcripts of discourse between co-located partners related to the quality of their collaboration and learning gains?

RQ2: Are Coh-Metrix indices different across experimental conditions?

RQ3: Are Coh-Metrix indices associated with MMLA measures (e.g., Joint visual attention, physiological synchrony, nonverbal behaviors) that were previously significantly correlated with collaboration quality?

RQ4: What Coh-Metrix indices are most meaningful for estimating a group's collaboration?

RQ5: Can Coh-Metrix indices be used to train supervised machine learning algorithms to predict collaboration quality?

4. METHODS

4.1 The Study

Participants with no self-reported prior programming or robotics knowledge ("novices") were paired randomly with an unknown partner and tasked with programming a robot to solve a series of increasingly complex mazes in 30 minutes. During the activity, mobile eye-trackers recorded participant gaze data, bracelets captured electrodermal activity, and a motion sensor collected movement and position data. A 2x2 study design was employed to test two different collaboration interventions: an informational intervention that described the benefits of collaborating on tasks and a visualization intervention that graphically plotted relative verbal contributions from each participant from the previous 30 seconds. The informational intervention is the primary on discussed here as the other did not result in significant differences in dependent measures. All participants gained knowledge of basic programming skills according to a pre-post survey ($t = 6.18$, $p < 0.001$) and a 7 percentage point increase in collaboration quality was associated with a 2 percentage point increase in code quality when controlling for gender and prior education ($p < 0.001$). For more details of experimental design and overall results, see Table 1 and [4]. Figure 1 shows a typical image of the experiment in progress.

4.2 Participants

Forty-two dyads completed the study and the first sessions each researcher conducted were removed to improve overall fidelity ($N = 40$ groups). Participants were drawn from an existing study pool at a university in New England in the United States. 62% of participants reported being students at the university of various levels, with ages ranging from 19 to 51 years old with a mean age

Table 1. Summary of measures from study.

Independent Measures	Process Data	Dependent Measures
Control Condition: no intervention Treatment Condition: informational intervention orally delivered by researcher prior to beginning of main portion of study.	Eye-tracking: Joint visual attention by dyads on different areas of interest, amounts of time looking at areas of interest Electrodermal activity: differences between individuals, synchrony measures, rates of change Movement and posture: proximity, alignment, bimanual coordination, total movement, leaning, synchrony measures	Expert ratings of collaboration Task performance measures Survey gains

Figure 1. Experimental setup from the study.

of 27 years. 60% of participants identified as female. Participants were paid $20 per 90-minute session of the study.

4.3 Procedure

Prior to the main activity, participants signed informed consent paperwork and took a 5-minute pre-survey pertaining to simple programming tasks. Once the survey was complete, sensors were applied to the participants and calibrated while the function of each sensor was explained.

Next, participants were shown a tutorial video explaining how to write code in Tinker, a block-based environment designed for use with the sensors and motors on the robot. Participants were then given 5 minutes to write a simple program to move the robot forward past a red line on the table in front of them. After the completion of this tutorial activity, participants were shown a second tutorial video that explained more advanced features of the programming environment like setting threshold sensor values for triggering commands and using conditional statements. A reference sheet was also provided to participants that summarized the content of both tutorial videos.

At this point, groups in the Intervention condition were read a summary of several research findings relevant to collaboration such as the importance of equity of speech time in high quality collaboration. Dyads in the Control group were given no such information after completing the tutorial (N = 20 groups in each condition).

Dyads then had 30 minutes to write code to guide their robot through a series of increasingly complex mazes. Participants did not know the arrangement of the mazes ahead of time and were prompted to write code that could solve any simple maze. Once the robot completed a maze twice successfully, the next one was provided by the researcher. During this portion of the activity, predetermined hints were provided to groups at 5-minute intervals to ensure common pitfalls identified in pilot resting were avoided. Following the completion of this portion, the post-survey was administered, demographic data was collected, sensors were removed, and participants were paid and debriefed.

4.4 Dependent Measures

Dyads' collaboration was evaluated live by the researcher conducting the session. Quality of collaboration was assessed on nine different scales derived from Meier, Spada, and Rummel's work on assessing collaboration in CSCL [16]: sustaining mutual understanding, dialogue management, information pooling, reaching consensus, task division, time management, technical coordination, reciprocal interaction, and individual task orientation. Each scale was on a -2 to 2 scale, and all scales were added together to generate an overall collaboration rating for dyads. Multiple researchers conducted sessions of the study and thus coded dyads' behavior. Researchers double coded 20% of the sessions from videos collected during the session and achieved an inter-rater reliability of $\alpha = 0.65$ (75% agreement).

Learning of computational skills (identifying a bug in block-based code, anticipating the output of a code segment, describing how to do a task with pseudocode, etc.) was assessed individually via pre- and post-tests with four questions each. These measures were adapted from [17, 18]. Researchers coded a subset of the responses to 100% agreement based on their demonstrations of understanding of computational thinking principles then coded the remaining surveys with the developed rubric.

4.5 Data Pre-Processing

Data collected by the eye-trackers, wristbands, and motion sensors each needed to be processed individually prior to merging for analysis. See Reilly, Ravenell, and Schneider for details on the processing of the Kinect motion sensor data [15] and Dich, Reilly, and Schneider for use of the electrodermal activity data from the Empatica E4 bracelet [19]. Four different physiological synchrony measures were calculated for movement and electrodermal (EDA) data: Signal Matching (SM), Instantaneous Derivative Matching (IDM), Directional Agreement (DA) and Pearson's Correlation (PC). SM was calculated as the differences in area between the plots of the team members' EDA, IDM calculates how closely the slopes of the physiological signal curves match, DA identifies whether individuals' signal data increase or decrease at the same time, and PC looks for a linear relationship between EDA data of both participants. For details on the calculation of these measures, see [19].

For use in this study, the eye-tracking data was summarized in two different ways: The proportion of time both participants were looking at the same spot (joint visual attention, see [20]) and the proportion of time spent looking at various areas of interest around the room (the computer screen, the maze, the robot, etc.) as determined by the fiducial markers placed on all objects in the experiment.

Audio recordings of sessions were transcribed using multiple iterations of assignments to Amazon Mechanical Turk workers until it was suitable for analysis. Transcripts were formatted via Python to match the requirements of Coh-Metrix and were sent to the Institute for Intelligent Systems at the University of Memphis for analysis. All analyses in this study were done in Python using the scikit-learn package [25].

5. RESULTS

5.1 RQ1: Correlations with Collaboration

Indices positively correlated with our rating of collaboration include more dialogue between partners (as measured by the number of sentences and words uttered), the use of adverbs, the CELEX word frequency (how often content words appear in sentences), and the familiarity of content words used. Deep cohesion (the use of causal connectives to signify causal relationships) and increased temporality (cues about temporality and tense) were also significantly positively correlated with collaboration. Additionally, the Coh-Metrix L2 readability score (use of simple grammar that an English language learner could

more easily parse) was also positively correlated. An appendix with definitions of all indices discussed here is provided at the end of this paper (also available in Appendix A of [3]). Significant correlations are listed in Table 2.

Table 2. Indices correlated with collaboration (red indicates negative correlation).

Index	Value (p-value)	Description
DESSC	0.51 (0.0087)	Descriptive indices that describe the number and length of paragraphs, sentences, and words
DESWC	0.735 (<0.0001)	
DESPL	0.51 (0.0087)	
DESWLsy	-0.27 (0.023)	
DESWLsyd	-0.36 (0.019)	
DESWLlt	-0.41 (0.0081)	
DESWLltd	-0.17 (0.016)	
PCNARz	0.53 (0.0042)	Text easability principal component scores
PCNARp	0.55 (0.0023)	
PCDCz	0.476 (0.039)	
PCTEMPz	0.53 (0.0068)	
PCTEMPp	0.52 (0.0064)	
PCCNCz	-0.41 (0.028)	
PCCNCp	-0.48 (0.0063)	
LDTTRc	-0.76 (<0.0001)	Lexical diversity (unique words per total number of words)
LDTTRa	-0.76 (>0.0001)	
LDVOCD	-0.25 (0.047)	
CNCCaus	0.45 (0.037)	Incidence of connectives
CNCLogic	0.49 (0.042)	
CNCADC	0.36 (0.036)	
SMINTEr	0.40 (0.044)	Ratio of intentional particles to intentional verbs
SYNNP	-0.43 (0.0012)	Number of modifiers per noun phrase, mean
DRPVAL	-0.33 (0.020)	Agentless passive voice density, incidence
WRDNOUN	-0.43 (0.0016)	Word information (part of speech category, syntactic categories)
WRDADV	0.47 (0.028)	
WRDPRP2	-0.68 (0.0005)	
WRDFRQc	0.43 (0.0004)	
WRDFRQa	0.16 (0.038)	
WRDAOAc	-0.28 (0.048)	
WRDFAMc	0.52 (0.0008)	
WRDCNCc	-0.46 (0.0063)	
WRDIMGc	-0.51 (0.007)	
WRDHYPv	-0.49 (0.0047)	
WRDHYPnv	-0.44 (0.0020)	
RDL2	0.36 (0.023)	Coh-Metrix L2 Readability

Indices negatively correlated with collaboration include the mean number of syllables per word, the number of nouns, the lexical diversity (unique number of total words), and hypernymy for nouns and verbs (using specific words instead of general ones.) Features that indicate the difficulty of understanding text are generally negatively correlated with collaboration, such as modifiers per noun phrase (complex syntax places higher demands on working memory), agentless passive voice, and the ratio of intentional particles to intentional verbs (a higher ration indicates more inference is needed to understand the text.) Indices for specificity of content words are also negatively correlated, such as concreteness and imageability (how easy it is to create a mental image of the word.)

5.2 RQ1: Correlations with Learning Gains

The magnitude of participant learning gains on the pre-post survey was also correlated significantly with several Coh-Metrix indices. Unlike collaboration, learning gain is positively correlated with lexical diversity ($r = 0.41$, $p = 0.041$) which indicates that use of specific language may aid learning of computer science principles. On the other hand, learning gain is negatively correlated with pronoun incidence ($r = -0.44$, $p = 0.041$) and referential cohesion ($r = -0.34$, $p = 0.049$) indicating that use of vague, overlapping language by the dyad is associated with lower gains on the survey.

5.3 RQ3: Differences by Condition

Differences between the Control and Intervention conditions could also be seen in their discourse during the activity. According to paired t-tests, groups in the Control condition had fewer words ($t = -3.4$, $p = 0.0019$), fewer adverbs ($t = -2.30$, $p = 0.029$), fewer sentences ($t = -2.17$, $p = 0.038$) and shorter paragraphs ($t = -2.17$, $p = 0.038$) than those in the Intervention condition. Additionally, the Control condition discourse had higher lexical diversity than the Intervention condition (LDTTRc: $t = 2.93$, $p = 0.0065$; LDTTRa: $t = 3.03$, $p = 0.0049$) which was shown above to be negatively correlated with collaboration.

With respect to the differences between groups that saw a visualization intervention that plotted relative verbal contributions from each participant, only the L2 Readability score differed significantly between conditions ($t = -2.16$, $p = 0.039$). As this intervention did not result in significant differences in our outcome measures, it makes sense that the impact on our Coh-Metrix indices is also minimal.

5.4 RQ4: Correlations with MMLA Values

We also explored whether any dyad-level features engineered from our sensor data might correlate with the Coh-Metrix indices that were previously seen to be significantly related to collaboration. Out of 30 features from our Kinect, eye-tracking, and EDA data, four were significantly correlated with three or more indices shown in Table 2. Correlation coefficients for these four features are shown in Figure 2.

From our eye-tracking data, the amount of time participants spent looking together at neither the maze nor the computer (aoi_0) was significantly negatively correlated with word length ($r = -0.42$, $p = 0.034$), lexical diversity ($r = -0.40$, $p = 0.042$), and syntactic complexity ($r = -0.39$, $p = 0.049$). Those three indices were also negatively correlated with collaboration. The amount of time participants spent looking at the maze and robot but not the computer (aoi_5) was positively associated with second person pronouns ($r = 0.65$, $p < 0.001$) and L2 readability ($r = 0.39$, $p = 0.047$) but negatively related to word length ($r = -0.42$, $p = 0.034$), word diversity ($r = -0.47$, $p = 0.016$), and use of the passive voice ($r = -0.40$, $p = 0.046$). This also follows a similar pattern to our correlations with collaboration.

The directional agreement (DA) of the dyad is calculated as the proportion of time where EDA for both participants was increasing or decreasing at the same time (in other words, it is a measure of

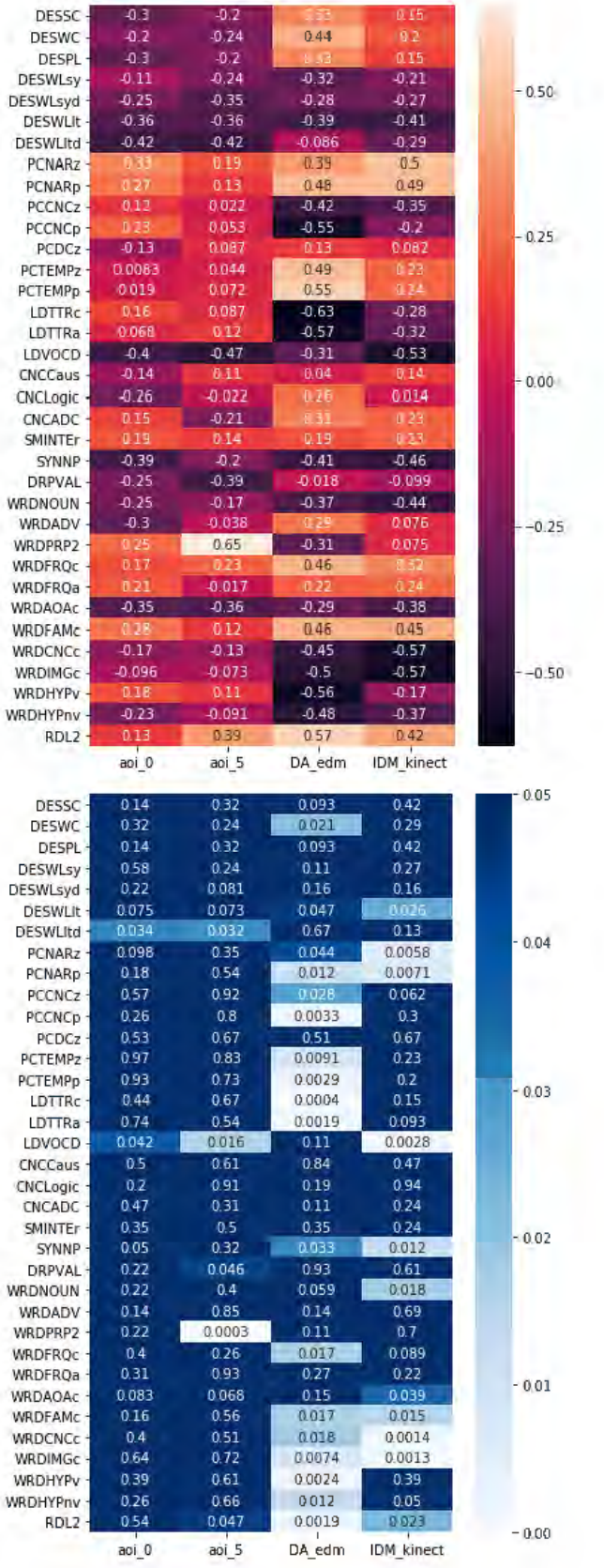

Figure 2. Correlation coefficients (top) and p-values (bottom) for select MMLA features.

physiological synchrony). Higher DA was positively associated with the number of words used (r = 0.44, p = 0.021), narrativity (r = 0.48, p = 0.012), temporality (r = 0.55, p = 0.003), word frequency (r = 0.46, p = 0.017), word familiarity (r = 0.45, p = 0.017), and L2 readability (r = 0.57, p = 0.002). DA was significantly negatively associated with lexical diversity (r = 0.44, p = 0.021), syntactic complexity (r = -0.41, p = 0.033), word concreteness (r = -0.45, p = 0.018), imageability of content words (r = -0.50, p = 0.007), and hypernymy (r = -0.56, p = 0.002). The directions of these correlations also fit with what we observed.

The Instantaneous Derivative Matching (IDM) of the Kinect movement data is calculated as the proportion of time where movement of both dyad members is either increasing or decreasing at a similar rate. IDM of movement was positively associated with narrativity (r = 0.50, p = 0.006), word familiarity (r = 0.45, p = 0.015), and L2 readability (r = 0.42, p = 0.023). Movement IDM was negatively correlated with word length (r = -0.41, p = 0.026), lexical diversity (r = -0.53, p = 0.003), syntactic complexity (r = -0.46, p = 0.012), the number of nouns used (r = -0.44, p = 0.018), word concreteness (r = -0.57, p = 0.001), imageability of content words (r = -0.57, p = 0.001), and hypernymy (r = -0.37, p = 0.049). Again, these correlations are of similar magnitude and direction as those seen in our results from collaboration.

5.5 RQ5: Predicting Collaboration

In order to explore how we might be able to use the Coh-Metrix indices to predict quality of collaboration, we classified dyads in terms of their collaboration ratings using a variety of typical machine learning classifiers. All 106 Coh-Metrix indices were used as features to classify the 40 groups. Missing values were imputed with their column means and all features were normalized prior to their use.

We first separated our participants into two groups based on the median value of group collaboration. We trained a Naïve Bayes classifier, a support vector machine (SVM), and a Random Forest (RF) model [21] on our entire data. NB usually works well with text data [26], SVM excels at binary classification [27], and RF along with other tree-based classifiers have been used successfully in the EDM community with a wide variety of educational datasets [28]. These algorithms were also selected as they are computationally rapid to implement once tuned and may be used in a real-time nature during a future intervention. The alpha, loss, and penalty used by SVM as well as the number of estimators, maximum depth, and criterion function for RF were selected by grid search with 5-fold cross-validation (CV). To address issues of overfitting with a small sample size, we report both training accuracy as well as the highest average accuracy achieved by our 5-fold cross-validation.

As shown in Table 2, the Random Forest model outperformed both Naïve Bayes and SVM on the binary classification median split task. While the 100% train accuracy of the RF is surely due to overfitting, the high CV accuracy. All algorithms were able to outperform random assignment by large margins. We next separated out participants into four groups based on the quartile values of group collaboration. Our three models were fit with the same procedure and once again RF outperformed both other algorithms on the training data. When looking at validation results, however, the RF and SVM classifiers performed identically. Simpler models may avoid the overfitting issues leading to the reported 100% training accuracy.

Table 2. Accuracy of classifiers.

	Median Split Train	Median Split CV	Quartile Split Train	Quartile Split CV
Naïve Bayes	0.88	0.74	0.81	0.51
SVM	0.88	0.75	0.84	**0.53**
Random Forest	**1.00**	**0.84**	**1.00**	**0.53**

To gain more insight into how these classifiers made their assignments, we investigated which features the RF model for the quartile split problem was ranking as most important for making assignments. Figure 3 plots the eleven most important features for our classification problem. Beyond that point, the feature importance rankings are too similar to derive insight. It is important to note that importance here is agnostic of whether these features correspond to good or bad collaboration; they are simply the most meaningful for deciding between them. Second person pronoun incidence (WRDPRP2) and lexical diversity (LDTTRc, as measured by the type-token ratio [22]) are the most important features. Word count (DESWC), sentence count (DESSC), incidence of negation expressions (DRNEG), text easability (PCDCp), and verb phrase density (DRVP) also rate highly. Many of these features were previously seen to be significantly correlated with collaboration, but verb phrase density and the incidence of negations appear here in our analyses for the first time.

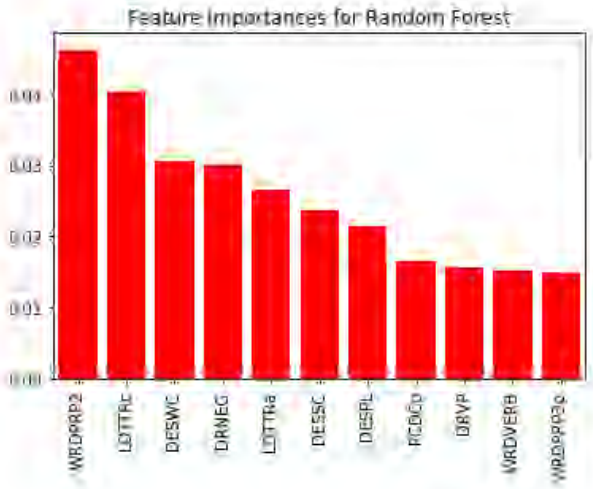

Figure 3. Feature importances for the quartile split Random Forest model.

6. DISCUSSION

In general, our findings indicate that our strongest collaborating dyads communicated more in terms of amount of words and sentences as well as the length of each utterance before the other participant would interject. In addition, these groups used more abstraction when referring to content words and terms and employed basic words and grammar to convey meaning is a direct fashion. They avoided using the passive voice or pronouns while reaching a consensus on a simple shared set of words to describe the task and their actions. While synonyms and extraneous modifiers were not used by strong collaborators, adverbs were used to define particular actions the robot needed to perform, and the use of logical, causal, and temporal connectives indicates a value to explicitly linking actions across space and time to meet the desired outcome. These indices of cohesion jointly allow collaborators to negotiate a shared problem space regardless of English language proficiency or level of education.

To ground the above findings, here is an example of a low collaborating dyad's discourse regarding programming the robot for a new maze:

A: So let's, we can do, no, yeah, we can put up, if yeah and if it's that then it goes this.
B: Then we add, turn right.
A: Yeah it will go right and then it will take for, wait for 10 seconds and then take a left. Also take a left.
B: Yeah, go, go forward. Go forward. Left, then we go straight.
A: Let's go forward. Left and then right. Then left and right.

In contrast, this is dialogue from a high collaborating group at a similar point in the activity:

C: So let's try changing this value to...greater than the second "If Do".
D: Okay. I just want to see if, oh, what did I do there, I just want to see if that what difference that makes.
C: Perfect. All right, are you ready?
D: Yep. Nope, all right.
C: Okay so, we've got it going forward and turning right so at least the right works. That one's correct now.
D: Now... if we change this number so let's go back to the widget.
C: Okay, okay. I think I've got it, so we needed to turn so when we got it turn right, we need to maybe check to if it turns left or right needs to be "greater than".

In the second dyad's discourse, more complete yet simple grammar and explicit markers of turn-taking result in a much fuller discourse that is easier to track. Their use of causal language and conditionals and implies a greater grasp of the content of the activity.

The importance and effects of cohesion are typically much higher for low-knowledge readers, with this relationship dubbed the "reverse cohesion effect" in discourse literature [23]. As none of our participants had any prior knowledge of robotics or computer programming, the importance of cohesion in participant discourse is likely to be crucial in similar educational settings. Reading skill and young age can also interact with this effect but these issues were not confounders in our study due to the use of oral dialogue and our population being solely adults.

As far as how the Treatment and Control groups differed, the Control groups typically communicated less (fewer sentences with shorter exchanges), used less adverbs, and had a higher lexical diversity (which was shown to be negatively correlated with collaboration). This difference shows that even simple verbal cues delivered as an intervention prior to an activity can have a positive association with collaboration by fostering more cohesive communication. The effect size of an informational intervention such as this on collaboration might be effectively used as a baseline when comparing more elaborate interventions in similar activities in the future. It is also reassuring to see the low effect on the Coh-Metrix indices from the visualization intervention condition that had no effect on collaboration. This validates our strategy of using

these indices without coding scheme to assess collaboration quality in a variety of different experimental conditions.

Comparing average learning gains on the pre-post survey to the Coh-Metrix indices is difficult for several reasons. First, the survey was done at the individual level and by only using the mean change we ignore when participants unequally learned during the task. Second, gains may be susceptible to ceiling effects where high gains are not seen due to high performance on the pre-test. Third, the dyads were instructed to program the robot to solve mazes and thus their conversation revolved around that task. The activity certainly utilized the computer science principles that were assessed in the survey, but the discussion was not as specific as a tutor dialogue regarding programming.

Despite these issues and challenges, several Coh-Metrix indices reveal what types of markers in the discourse can signal learning taking place. While lexical diversity was negatively associated with collaboration, it appears to be beneficial for learning. Knowing and applying more terms for phenomena or problem-solving strategies may aid participants transfer their knowledge from the experimental task to the post-survey. Additionally, too much referential cohesion may make ideas difficult to separate out of context and thus more challenging to use in isolation on test questions.

It is worth noting that features engineered from all three of our MMLA sensors provided insight into how to assess collaboration using the Coh-Metrix indices identified as significantly related to collaboration. When joint visual attention fell outside of the tabletop or laptop (aoi_0), this could generally be interpreted as participants looking at each other (as relatively little time was spent with both participants simultaneously looking at the facilitator, the same spot on the wall, or anything unrelated to the task). The proportion of time spent doing this negatively correlated with word length, lexical diversity, and syntactic complexity (all of which are markers of poor collaboration). Eye contact is positively associated with problem solving and facilitates conceptual understanding in group settings [24] so this result triangulates established literature findings. Joint visual attention being more focused on the maze and robot instead of the laptop (aoi_5) correlated positively with second person pronoun use and readability while negatively correlating with word length, word diversity, and use of the passive voice. This can be interpreted as dyads communicating with each other more effectively by looking at the physical problem space and talking through the steps needed to solve the problem versus spending more time in the programming interface editing code.

Two of our measures of synchrony (directional agreement for EDA and instantaneous derivative matching for motion) are positively associated with indices deemed good for collaboration and negatively correlated with indices seen to be negatively related to collaboration. By focusing on these four features from our multimodal data, we might be able to automatically assess collaboration during trials, which could be used to provide formative feedback and design new interventions based on these measurements.

Finally, we used supervised machine learning algorithms in hopes of being able to detect and intervene while dyads are working together. The relative feature importances from our Random Forest classifier also shed light on what indices are most useful for assessing collaboration. Second person pronoun incidence, number of words, and lexical diversity appeared in our correlations with collaboration, while verb phrase density and the incidence of negations did not appear in our previous results. The emergence of negation in this model will need to be studied more thoroughly. Increased incidence of negation expressions may signal discord between the participants that could hinder the joint construction of meaning en route to problem solving. It is possible that the relationships between these features and collaboration are nonlinear and are thus not detected as readily by simple correlational analyses. Additionally, the overfitting of the models may be due to the lack of regularization of model complexity. With the range of -2 to 2 for the collaborative scoring, it might be more appropriate to fit a regression model instead of classifying the scores.

This preliminary work has several limitations that must temper the results. The small sample size of 40 groups leads to overfitting of our classifiers and may interfere with the ability of some of the Coh-Metrix algorithms to function. The designers recommend using a carpus of roughly 300 texts of 300 words each to study text easibility [3]. While the length of our transcripts exceeds these recommendations, it is unclear what effect our sample may have on this novel use of Coh-Metrix. We collected no reading level demographic data on our participants, nor did we ask for whom was English a first language. Additionally, this preliminary work does not address the important issue of how does communication (and thus collaboration) differ when participants have very different expressive language capabilities.

The developers of Coh-Metrix intended these indices to be the "low-hanging fruit" of computational linguistics, choosing to use simple metrics instead of complex computational linguistic models [3]. While this will likely be valuable for developing real-time dynamic interventions that can't be slowed down by computationally expensive operations, we need to compare these results to more complex models and other natural language processing methods. Future work will also explore "driver-passenger" models that investigate emergent leadership behavior and uneven talk time in the discourse as well as the role of eye contact in the quality of collaboration.

7. CONCLUSION

This research paves the way for real-time detection of (un)productive interactions from multimodal data, potentially facilitating the development of fail-soft real-time interventions to support collaborative learning. While Coh-Metrix is only available currently as an online service, similar analytical platforms can be run locally [9] and could offer advice based on specific issues detected in the discourse rather than general distribution of talk time. These indices of cohesion and easibility have proven to be versatile and serve as effective features for estimating the rough quality of dyadic discourse with regard to collaboration quality.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1748093. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] Roschelle, J. and Teasley, S.D., 1995. The construction of shared knowledge in collaborative problem solving.

In *Computer Supported Collaborative Learning* (pp. 69-97). Springer, Berlin, Heidelberg.

[2] National Research Council. 2012. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press.

[3] McNamara, D.S., Graesser, A.C., McCarthy, P.M. and Cai, Z., 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

[4] Starr, E., Reilly, J., and Schneider, B. 2018. Using Multi-Modal Learning Analytics to Support and Measure Collaboration in Co-Located Dyads. In *Proceedings of the 13th International Conference on the Learning Sciences*, 448–455.

[5] Dillenbourg, P., Järvelä, S. and Fischer, F., 2009. The evolution of research on computer-supported collaborative learning. In *Technology-Enhanced Learning*, 3-19. Springer, Dordrecht.

[6] Dillenbourg, P., Baker, M.J., Blaye, A. and O'Malley, C., 1995. The evolution of research on collaborative learning. Spada, E. and Reiman, P. Learning in Humans and Machine: Towards an interdisciplinary learning science., Elsevier, Oxford, 189-211.

[7] Blikstein, P. and Worsley, M., 2016. Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220-238.

[8] Graesser, A.C., McNamara, D.S., Louwerse, M.M. and Cai, Z., 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

[9] Crossley, S.A., Kyle, K. and McNamara, D.S., 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, *48*(4), 1227-1237.

[10] Duran, N.D., Hall, C., McCarthy, P.M. and McNamara, D.S., 2010. The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics*, *31*(3), 439-462.

[11] Graesser, A.C., Jeon, M., Yan, Y. and Cai, Z., 2007. Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, *15*(3), 199-213.

[12] D'Mello, S.K., Dowell, N. and Graesser, A.C., 2009, July. Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 9-16.

[13] Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M. and Siemens, G., 2016, April. Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 15-24.

[14] Schneider, B. and Pea, R. 2015. Does seeing one another's gaze affect group dialogue? A computational approach. *Journal of Learning Analytics*, *2*(2),107-133.

[15] Reilly, J., Ravenell, M., and Schneider, B. 2018. Assessing Collaboration Using Motion Sensors and Multi-Modal Learning Analytics. In K.E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th International Conference on Educational Data Mining*, 24 –257.

[16] Meier, A., Spada, H., and Rummel, N. 2007. A rating scheme for assessing the quality of computer-supported collaboration processes. *Computer Supported Learning*, 2, 63–86.

[17] Brennan, K. and Resnick, M. 2012. New frameworks for studying and assessing the development of computational thinking. Presented at the Proceedings of the 2012 annual meeting of the American Educational Research Association, Vancouver, Canada.

[18] Weintrop, D. and Wilensky, U. 2015. Using commutative assessments to compare conceptual understanding in blocks-based and text-based programs. Presented at the 11th Annual ACM Conference on International Computing Education Research.

[19] Dich, Y., Reilly, J., and Schneider, B. 2018. Using Physiological Synchrony as an Indicator of Collaboration Quality, Task Performance and Learning. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*, 98 – 110.

[20] Richardson, D.C. and Dale, R., 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*(6), 1045-1060.

[21] Breiman, L., 2001. Random forests. *Machine Learning*, *45*(1), 5-32.

[22] Templin, M. 1957. Certain language skills in children: Their development and interrelationships. The University of Minnesota Press, Minneapolis.

[23] O'Reilly, T. and McNamara, D.S., 2007. Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse processes*, *43*(2), 121-152.

[24] Joiner, R., Scanlon, E., O'Shea, T., Smith, R.B. and Blake, C., 2002, January. Evidence from a series of experiments on video-mediated Collaboration: Does Eye Contact Matter?. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*, 371-378.

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

[26] Rennie, J.D., Shih, L., Teevan, J. and Karger, D.R. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning*, 616-623.

[27] Scholkopf, B. and Smola, A.J. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[28] Romero, C. and Ventura, S., 2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

[29] Rummel, N. and Spada, H. 2005. Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences*, *14*(2), 201-241.

[30] Dillenbourg, P., Järvelä, S., and Fischer, F. 2009. The evolution of research on computer-supported collaborative learning. In *Technology-enhanced learning*, 3-19. Springer, Dordrecht.

[31] Ochoa, X. and Worsley, M. 2016. Augmenting Learning Analytics with Multimodal Sensory Data. *Journal of Learning Analytics*, *3*(2), 213-219.

10. Appendix: Coh-Metrix Indices

Index	Description	Index	Description
DESSC	Sentence count, number of sentences	CNCLogic	Logical connectives incidence
DESWC	Word count, number of words	CNCADC	Adversative and contrastive connectives incidence
DESPL	Paragraph length, number of sentences, mean	SMINTEr	Ratio of intentional particles to intentional verbs
DESWLsy	Word length, number of syllables, mean	SYNNP	Number of modifiers per noun phrase, mean
DESWLsyd	Word length, number of syllables, standard deviation	DRVP	Verb phrase density, incidence
DESWLlt	Word length, number of letters, mean	DRNEG	Negation density, incidence
DESWLltd	Word length, number of letters, standard deviation	DRPVAL	Agentless passive voice density, incidence
PCNARz	Text Easability PC Narrativity, z score	WRDNOUN	Noun incidence
PCNARp	Text Easability PC Narrativity, percentile	WRDADV	Adverb incidence
PCDCz	Text Easability PC Deep cohesion, z score	WRDPRP2	Second person pronoun incidence
PCTEMPz	Text Easability PC Temporality, z score	WRDFRQc	CELEX word frequency for content words, mean
PCTEMPp	Text Easability PC Temporality, percentile	WRDFRQa	CELEX Log frequency for all words, mean
PCCNCz	Text Easability PC Word concreteness, z score	WRDAOAc	Age of acquisition for content words, mean
PCCNCp	Text Easability PC Word concreteness, percentile	WRDFAMc	Familiarity for content words, mean
LDTTRc	Lexical diversity, type-token ratio, content word lemmas	WRDCNCc	Concreteness for content words, mean
LDTTRa	Lexical diversity, type-token ratio, all words	WRDHYPv	Hypernymy for verbs, mean
LDVOCD	Lexical diversity, VOCD, all words	WRDHYPnv	Hypernymy for nouns and verbs, mean
CNCCaus	Causal connectives incidence	RDL2	Coh-Metrix L2 Readability